DOCUMENT RESUME

ED 178 540 SP 015 077

AUTHOR Anderson, Linda M.
TITLE Classroom-Based Experimental Studies of Teaching Effectiveness in Elementary Schools.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Institute of Education (DHEW), Washington, D.C.
REPORT NO UTR&D-R-4102
PUB DATE 79
CONTRACT OB-NIE-G-78-0216
NOTE 53p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Achievement; Class Management; Classroom Environment; *Classroom Observation Techniques; *Educational Research; Effective Teaching; Evaluation Methods; *Research Methodology; *Teacher Behavior

ABSTRACT

Three recent large-scale experimental studies have built on a data base established through several correlational studies of teaching effectiveness in elementary school. These three studies have in common a treatment that addresses several routine teaching tasks and suggests some principles and techniques for effectively fulfilling those tasks. All three studies yielded results indicating that the treatment teachers did use many of the behaviors suggested to them and that their students had higher adjusted achievement scores. In this article, the three studies are reviewed and suggestions are made about future experimental studies of teaching effectiveness. (Author/JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Classroom-based Experimental Studies of Teaching Effectiveness in Elementary Schools

Linda M. Anderson

The Research and Development Center for Teacher Education

The University of Texas at Austin

R&D Report No. 4102

SP015077

This paper was produced within the Correlates of Effective Teaching Program, Carolyn M. Evertson, Director, and supported in part by the National Institute of Education Contract OB-NIE-G-78-0216, Correlates of Effective Teaching Program, The Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred.

Footnotes

The author wishes to thank Charles W. Anderson, Jere E. Brophy, and Carolyn M. Evertson for comments and suggestions about the paper, and Carol Culp for manuscript preparation.

Abstract

Three recent large-scale experimental studies have built on a data base established through several correlational studies of teaching effectiveness in elementary schools. These three studies have in common a treatment that addresses several routine teaching tasks and suggests some principles and techniques for effectively fulfilling those tasks. All three studies yielded results indicating that the treatment teachers did use many of the behaviors suggested to them and their students had higher adjusted achievement scores. In two of the studies, the process-product correlations suported earlier work. In this article, the three studies are reviewed and suggestions are made about future experimental studies of teaching effectiveness. Those suggestions are organized around the topics of creating treatments that maximize the chances for ecologically valid results, and of designing studies and choosing measures that allow as much experimental control and monitoring of the natural setting as possible.

During the last ten years, there has been much research that systematically relates teaching behaviors and classroom characteristics to student learning. There was once pervasive pessimism regarding the practical efficacy of research on teaching effectiveness, and even widespread doubt that teachers have any important effects on students or that any teacher effects could be systematically documented (e.g., Heath & Nielson, 1974). However, recent reviewers agree that a knowledge base has developed, and a cohesive (although perhaps not inclusive) picture can be drawn of "effective teaching," at least for educational settings and outcomes that have been studied most intensively (Borich, 1977; Brophy, in press; Brophy & Evertson, 1976; Good, 1979; Good, Biddle, & Brophy, 1975; Rosenshine, 1976; Medley, Note 1).

Most of the work discussed in these references has been correlational. However, there have recently appeared three large-scale classroom-based experimental studies of teaching effectiveness in the elementary grades (Anderson, Evertson, & Brophy, 1979; Good & Grouws, in press; Program on Teaching Effectiveness, Note 2). These three studies represent an important set of related findings and an approach to research on teaching that adds to and extends the existing large body of correlational process-outcome studies.

The purpose of this article is to discuss the application of experimental design to the study of classrooms. Three major studies completed at the elementary level are reviewed,[1] and recommendations for future work are based on them and other sources.

Correlational Studies of Teaching Effectiveness

The three experimental studies grew out of a knowledge base generated by several years of research that can be characterized in these ways:

1) Data were collected in classrooms where teachers and students were moving through their typical routines, without intervention by the researcher.

That is, the studies were naturalistic, and the observers and investigators did not play important roles in determining classroom processes. Therefore, the data related natural occurrences throughout the year to the outcomes of interest, usually student learning as measured on achievement tests. In order to get as accurate a picture as possible, the data were collected over several observations. Because of this focus on natural occurrences, the relationships between classroom processes and student outcomes were correlational.

2) Most of the variables used to describe process-outcome relationships represented "typical" activities and tasks of elementary school teaching. Rather than focusing on special or new curricula, instructional programs, or organizational patterns, the studies focused on common teaching activities that occur in a variety of settings. Examples of these are: arranging students for instruction, questioning students and providing feedback to their answers, allocation of time, maintenance of order and monitoring of student behavior, providing for classroom routines and procedures, presenting information, and assigning work to students.

3) The purpose of most of the studies was to describe "effective teaching," with "effective" defined as various criteria, usually student achievement in the basic skill areas. Therefore, the purpose was not only to describe classroom processes, but to learn about relationships between those processes and the outcome of interest. Such research has been called "improvement-oriented," implying a practical concern with benefiting the educational process (Koehler, Note 4). Because of this goal, researchers conducting studies of teaching effectiveness have typically been concerned about the generalizability and application of their findings to the "real world," and have assumed that the first step in the development of effective teachers was to define effective teaching.

Therefore, correlational studies have generally been perceived as an important step toward the longer-range goal of understanding and prescribing teaching practices effective for various objectives. In and of themselves, these correlational studies have serious limitations when one's avowed purpose is to improve educational practice. The most serious limitations are these:

1. Because the relationships between teacher behaviors and student outcomes are correlational, it cannot be determined that the teachers' actions actually caused the outcome. Therefore, it would be inappropriate to assume that all teacher behaviors that are highly correlated with student achievement actually preceded and caused that achievement. Statements about what teachers should do cannot be derived directly from correlational data.

2. Even if a causal relationship could be assumed (and in some cases it is very likely to exist), this does not mean that the information can be translated easily into improved practice by teachers who do not already possess those "effective" skills. Therefore, an important research question is whether or not the identification of effective teaching behaviors can lead to improved instruction, and if so, what are the most effective means of providing the information.

Experimental Studies of Teaching Effectiveness

The classroom-based experimental studies of teaching effectiveness are targetted to these questions about the causal nature of teacher effects and the utility of research-based descriptions of effective teaching. In such studies, teachers are encouraged to perform specific behaviors that have been associated with a desirable outcome, and data are collected on teachers' behaviors and associated effects on students. Such an experimental paradigm was recommended by Rosenshine and Furst (1973), who described a cycle of research on teaching that developed from descriptive to correlational to

experimental studies. Dunkin and Biddle (1974) also recommended that researchers pursue experimental studies to verify earlier findings.

The experimental paradigm is certainly not new to educational research, or even to research on teaching effectiveness. However, with the exceptions described in the next section, most of the experimental work has relied on laboratory-like settings or artificial controls within classrooms in order to study a small number of variables over a relatively short period of time. For example, teachers were given scripts to follow, so that different levels of questioning and feedback behavior could be systematically varied and related to test scores for an ecology unit (Program on Teaching Effectiveness, Note 5). In such studies, the demands on teachers were greatly reduced from those of the classroom, and the outcome measures were very specific to the tasks in the study. Therefore, any relationships between teaching behavior and outcome were not easily generalizable to classroom teachers, who must deal with many complex, simultaneous demands without the benefit of a script and limited responsibilities.

Because of the highly controlled nature of such experiments, they are often viewed as clean but trivial, at least when evaluated for their immediate relevance to teaching practice. In order to conduct experiments that verify earlier findings but also retain external validity (i.e., "real-world" relevance), it has been necessary to move into the classroom and conduct studies that blend experimental control with realistic interventions.

Considerations for Experimental Design in Classroom Studies

Much has been written about various ways of examining validity in an experiment. Campbell and Stanley's (1963) classic piece primarily emphasized internal validity--the degree to which one can isolate factors that contribute to a causal relationship. Bracht and Glass (1968) extended Campbell and

Stanley's briefer discussion of external validity (the extent to which experimental results can be generalized beyond the experimental setting.) They described two important kinds of external validity: population validity (the degree to which generalizability is limited to the population used in the experiment), and ecological validity (the degree to which generalizability is limited to the setting used in the study). They suggested several ways in which any experimental setting may contribute to effects beyond those caused by a treatment, such as Hawthorne effects, pretest and posttest sensitization, and novelty and disruption effects. In applying this concept to the study of teaching and classrooms, one must ask if the act of "treating" any elements of the setting changes the setting sufficiently to prevent generalization to other classrooms and teachers. That is, will the same treatment, outside of an experimental study, have the same effects?

Shulman (1970) suggested that educational researchers should also be concerned with "task validity" as another source of external validity. That is, are the tasks (or mental operations) performed by subjects in an experiment like those utilized in the larger setting of interest? He expanded on the idea of task validity by suggesting that researchers must find better ways to study and characterize learning environments, recognizing their inherent complexities.

Snow (1974) reviewed earlier work and suggested that an additional concern should be that experiments be representative. He stated that the "biggest threat to external validity may come when the experiment does not fit the nature of the behavior being studied and, furthermore, does not include the means of discovering this fact." (p. 265) Snow drew heaviliy on the work of Brunswik (1956), who emphasized the adaptive, active nature of psychological processes, and suggested that the experimenter should adapt

methodology to fit this form of phenomenon rather than trying to examine it through simpler methodologies that isolate and decontextualize discrete behaviors. Snow suggested several ways in which representative and quasi-representative experiments could be conducted, including moving into applied settings such as classrooms, conducting intra-experiment observations to document actual processes, extending the life of treatments, and sampling several experimental effects. Many of Snow's suggestions were carried out in the three studies reviewed below, and in these respects, they differ from more traditional educational experiments.

Cook and Campbell (1976) discussed the design and conduct of experimental studies in field settings. Although they focused on examples drawn from industrial and organizational psychology, their points are also very applicable to research in schools. They refined the meaning of external validity to distinguish between generalizability of results to other settings and populations and "construct validity." The latter refers to the accuracy with which the causal mechanisms and effects have been described. That is, when a treatment effect has been detected, what aspect of the treatment was responsible? In order to assess construct validity, one must understand how the treatment, as viewed by the experimenter, may be confounded with other causal mechanisms. As examples, they describe Hawthorne effects (increased productivity was related to the extra concern and attention that came with membership in the treatment groups) and experimenters' expectancies communicated to treatment subjects that affect their responses. Such effects may be difficult to separate from the content of the treatment--the constructs that are supposedly being examined in a controlled manner.

Construct validity is an important concern in classroom research. As will be seen in reviewing the three studies, it is often difficult to

establish what aspects of the treatment content and process most account for an overall treatment effect. Such questions must be resolved if one is to generalize about the effects of the treatment to other settings.

In summary, thinking about educational experiments since Campbell and Stanley (1963) has reflected increasing concern with external validity (including construct validity) and a recognition that natural (i.e., representative) settings often have to be utilized if there is to be generalizability. This imposes two difficult tasks on the experimenter. First, a greater concern than before about external validity does not reduce the need for internal validity, even though experimental control is more difficult to exert in "real-world" settings. Second, the experimenter must have an adequate understanding of the complexity of the natural setting in order to select treatments that are relevant (i.e., match the demands of that setting) and to select the potentially most important intervening variables to measure. In short, the experimenter must balance rigor and relevance.

Research in any field setting poses difficulties, and the classroom is no exception. At any given moment, the teacher must be aware of and coordinate the activities of 20 to 35 students in order to move toward a variety of goals established by the teacher, the school administration, and society. Classroom ecology has been characterized as including multidimensionality, simultaneity of events, immediacy of demands, unpredictability of interruptions, and a history that colors the participants' perceptions (Doyle, 1977).

In spite of this complexity, classroom researchers have been able to identify several characteristics of effective teaching, especially in the primary grades (e.g., Brophy & Evertson, 1976; Good, 1979; Medley, Note 1). The result of such research has not been a list of discrete teacher competencies that can be considered in isolation from one another. Instead,

the effective teacher has been portrayed as orchestrating and integrating many discrete skills. That is, an effective teacher does many things well. For any study, correlational or experimental, to contribute to knowledge about classrooms, this fact must be recognized.

Given the incredible complexity of the classroom and the many influences on its processes and outcomes, the experimenter's task of juggling rigor and relevance in a classroom-based study is an especially difficult one. However, the most important similarity of the three studies reviewed below is a common pattern of results that indicates high payoff. In each case, providing teachers with integrated research-based principles of instruction preceded changes in teacher behaviors (in the desired direction) and an increase in student achievement. The implication of these results is that a classroom-based, experimental approach that incorporates several aspects of instruction holds promise for the study of teaching effectiveness in spite of its inherent difficulties.

Summaries of the Three Studies

Each of the three studies included a treatment given to one or two experimental groups of elementary teachers. The treatment in each case contained several interrelated suggestions about instruction. All suggestions had a basis in earlier research or experience, and most were taken from large-scale correlational studies of teaching effectiveness. The treatments focused on teaching methods, rather than curriculum content.

The studies were conducted in natural classroom settings, with teachers using the activities, materials, and standard curricula that they would have used in the absence of an experiment. The experimenters did nothing during actual instruction except to observe on a regular basis to gather data on

treatment implementation. The data were collected over a large portion of the school year, using complex coding instruments.

The content of the three treatments differs because the studies focused on different aspects of instruction. However, each may be said to reflect the general principles of "direct instruction" as described by Rosenshine (1976). That is, there was an emphasis in each treatment on teacher leadership in instruction, making sure that all students have adequate opportunities to be exposed to instruction and to practice basic skills.

Three important questions were addressed by the data in each study:

1. Did the teachers in the treatment group use the suggestions given in the treatment?

2. Were there differences in the achievement of the students who had treatment teachers and those who were instructed by control teachers?

3. Were the process-outcome relationships within the studies those expected on the basis of earlier studies? (That is, did the teacher behaviors predict achievement as in earlier correlational studies?)

The Missouri Mathematics Effectiveness Project (Good & Grouws, in press)

Background. A instructional model was designed for use by fourth-grade teachers in conducting mathematics lessons. Most of the suggestions in the math lesson model were based on an earlier correlational study (Good & Grouws, 1977), in which teachers who were consistently more or less effective in producing mathematics achievement were selected for observation. Data about their behavior were integrated with other recent research on mathematics teaching, yielding the lesson model that served as the treatment in the experimental study.

Treatment content. The model was a system of instruction that focused on student comprehension, systematic preparation of students for each stage of

the lesson, distributed practice, and explanations and presentations by the teacher. Specifically, teachers were asked to conduct whole-class lessons according to a suggested routine. Regular times for review were scheduled, the development portion of the lesson was emphasized, and teachers were urged to pace the lesson to maintain student attention and involvement and to also monitor student performance. Student accountability for seatwork and homework was emphasized. Specific suggestions were made for accomplishing each step.

Treatment administration. Treatment teachers attended a training session and were given copies of the 45-page manual about the lesson model, which contained definitions, rationales, and detailed descriptions of lesson components. Two weeks after the treatment began, the experimenters met again with the treatment teachers to answer questions. There was no further training.

Subjects. The sample included 40 volunteer teachers from 27 predominantly lower-SES schools in an urban district. All were fourth-grade teachers who taught math through a semidepartmentalized plan.

Design. Schools were randomly assigned to experimental conditions after being matched for SES-level. (Assignment was done by schools to prevent teachers in the same school from being assigned to different conditions, and therefore possibly spreading treatment information to the control classes.) The treatment group contained 21 teachers, and the control group (termed the "delayed-treatment" group) included 19 teachers. With a few exceptions, each teacher was observed six times.

Teachers in the delayed-treatment group were aware of the purpose of the study. The experimenters encouraged them to do their best by telling them that their mathematics teaching was under scrutiny, that their students' achievement would be evaluated, and that after the study, they would receive

feedback on their own teaching as well as the instructional model. (Therefore, their treatment was to be delayed, but not denied to them, and so they were not like typical control groups.) The experimenters reasoned that creation of a strong Hawthorne effect in both groups through encouragement and attention would enable them to determine if effects of treatment content added to the effects of enhanced motivation. Hence, both groups of teachers were made to feel accountable and were given extra attention, but only the immediate-treatment teachers had received the lesson model when observations were made. Delayed-treatment teachers were encouraged to teach as they normally did.

Measures. The primary classroom process measures were those used in the earlier correlational study, plus a summary checklist used to estimate implementation of treatment behaviors. The students' math achievement was tested before and after the treatment period (October through January) by SRA math subtests. In addition, a test designed to measure the content covered by all teachers was given in January.

Results: implementation of treatment. Analyses of observation data revealed that the treatment was fairly well implemented by most of the immediate-treatment teachers, although there was less implementation of the principles on the development portions of the lesson. However, on all general implementation measures, immediate-treatment teachers were performing more in line with the lesson model than delayed-treatment teachers.

The delayed-treatment teachers reported that they had given more thought to teaching math that year, but they had not significantly altered their regular teaching practices.

Results: effects on achievement. The performance of the students in the immediate-treatment group exceeded the performance of the delayed-treatment

group on all measures, in spite of the fact that the former group's entering scores were significantly lower than the latter group's. The delayed-treatment group's average percentile also rose during the period of the study, although the gains were not as large as the immediate-treatment group's. The gains of the delayed-treatment group suggest that the encouragement and attention given to them had beneficial effects. However, their students did not gain as much as students whose teachers had been exposed to the instructional strategies in the lesson model.

Results: process-outcome relationships. Several components of the lesson model were correlated with residual gains. There were significant positive relationships with achievement for use of review, assigning homework, providing practice in mental computations, and requiring accountability for seatwork. There were near-significant positive results for suggestions about conducting seatwork. In addition, all of the general implementation measures were correlated positively with achievement gains.

Conclusions. Significant increases in student achievement in mathematics were related to teacher behaviors that were apparently influenced by the treatment given to the teachers.

The Stanford Experiment on Teacher Effectiveness (Program on Teaching Effectiveness, Note 2).

Source of treatment. The treatment in this study was based on 125 variables derived from four correlational studies of teaching effectiveness: Brophy and Evertson (Note 6), McDonald and Elias (Note 7), Soar (Note 8), and Stallings and Kaskowitz (Note 9). Each of these studies had in common a focus on teacher and student behaviors in elementary classrooms, with end-of-year reading achievement included as one criteria of effectiveness. The

investigators selected variables that significantly correlated with reading achievement at the second or third grade level.

Treatment content. The resulting treatment was a five-part program including 22 recommendations in three broad areas: behavior management and classroom discipline; instructional methods; and questioning and feedback strategies. Under the heading of behavior management, some suggestions were to avoid target and timing errors in discipline, and therefore to remain "withit"; to establish a system of procedures for students' personal needs; and to move around the room to monitor student behavior. Suggestions grouped under instructional methods included minimizing direction giving and miminizing time spent in organizing for instruction and in instructing very small groups and individuals. The purpose was to maximize the total amount of time that individual students were under the direct supervision of the teacher. Many of the suggestions under questioning and feedback strategies distinguished between appropriate teaching practices for students who were more or less academically oriented. The strategies discussed here included responding to incorrect answers and adjusting difficulty levels of public questions.

Treatment administration. Teachers in the treatment group received one manual a week describing the treatment for a five-week period. One treatment group (minimum-treatment) received the materials through the mail, while another group (maximal-treatment) received the same materials and also attended a two-hour in-service meeting each week to discuss the topic. The control group received no training. Each week, treatment teachers completed quizzes and questionnaires after reading the materials. During the spring semester, there was a "refresher course" given to the treatment teachers,

during which the minimum-treatment teachers received written materials, and the maximal-treatment teachers were videotaped and given personal feedback.

Subjects. Included in the study were 33 third-grade classes in 24 elementary schools in two adjacent school districts that were slightly above the state average SES level. All teachers in the study were volunteers who had been told the full design and rationale of the study before assignment to treatment groups. All teachers had at least four years of teaching experience. In 29 of the 33 classes, there was also a state program being implemented with an emphasis on parent involvement and individualization. This is important to note, since some of the treatment teachers reported that the suggestions made in the experimental treatment conflicted with the expectations of the state program, especially regarding use of individualized and small-group instructional modes (Mitman, Note 10).

Design. There were three levels of teaching effectiveness treatment: maximal, minimum, and control. In addition, there were two levels of participation (present or absent) in a parent-involvement program, resulting in a 2 x 3 design with five or six classes in each cell. The groups were formed by stratified random assignment within two districts, based on the mean class reading pretest scorees. Therefore, the treatment groups were comparable in terms of entering achievement level of the students. Each class was observed four times before the treatment period, five times during the five weeks of treatment administration (November through December), and seven times after the treatment had been given. Each observation lasted a full day.

Measures. The primary process measure was the Classroom Observation Instrument (Stallings & Kaskowitz, Note 9) with some modifications. During January and May additional high inference ratings were taken. Students were

tested in September, January, and May on a variety of aptitude, achievement, and attitude measures. Teachers completed tests of verbal fluency and a questionnaire regarding teaching style preferences before the treatment was given.

Results: implementation. In general, the two treatment groups had higher implementation scores than the control group; the minimum-treatment group's scores were higher than the maximal-treatment group's. On the overall implementation measure, the difference was significant ($p = .02$); separate analysis of three treatment components yielded less significant differences, but the results still favored the treatment groups. When the separate principles of the treatment were examined, the treatment groups had significantly ($p < .10$) higher implementation scores than the control group on five of 16 principles analyzed at the class level. (Results for other variables were not significant; however, trends in the expected direction were found for five of the remaining nine variables.)

Of special interest was the relationship between implementation and the teachers' pretest measures of verbal fluency and self-ratings of structuredness: There were significant positive correlations for each of the teacher measures and implementation. Unfortunately, means on both teacher measures were higher in the minimum-treatment group, despite the random assignment of classes to treatment groups. When the composite implementation scores were adjusted for verbal fluency and pretreatment structuredness, the main effect of treatment on implementation was no longer significant ($p = .163$).

Results: student achievement. Analyses of student achievement at the end of the school year revealed several interactions between the two programs (the teacher effectiveness program discussed here, and the parent involvement

program) and school district membership. The district effect was to reverse the order of the minimum and maximal groups, but in each case the treatment groups had higher adjusted achievement on reading tests. Classes receiving both teacher effectiveness training and the parent-assisted program had the highest adjusted reading achievement of all groups.

Results: process-outcome relationships. When using the composite measures of implementation, there were consistent although not strong positive relationships with adjusted achievement. However, only two out of nine correlations were significant at $p < .05$. This pattern of positive correlations was more evident for vocabulary scores than other outcome measures, and more evident for the questioning and feedback principles than for the other two parts of the treatment (Crawford & Stallings, Note 11). Stayrook and Crawford (Note 12) report the direction and strength of correlations of the individual process variables with achievement. About half of the correlations agreed in sign with the original studies, but only one variable out of 48 was significantly associated with Total Reading achievement at the level of $p = .05$, and it was correlated in the opposite direction expected.

However, the investigators noted that one difficulty in evaluating their results was that many of the variables (especially those derived from the Brophy and Evertson study) were not adequately measured with their observational instrument. The lack of clear process-outcome relationships may be a function of this.

Conclusion. Generally, the experiment fulfilled its original objective: teachers in the treatment groups used the suggested behaviors more than teachers in the control groups, and there was a corresponding difference in adjusted achievement favoring the treatment groups. However, the

process-outcome relationships do not strongly support the conclusion that the content of the treatment was most responsible for these changes in behaviors and outcomes. It is possible that simply participating in a treatment program and focusing attention on the goal of increased reading achievement could account for improved learning.

The Texas First-grade Reading Group Study (Anderson, Evertson, & Brophy, 1979)

Source of treatment. The treatment was an instructional model that presented information about small-group management to first-grade teachers to use in reading lessons. The primary source of the treatment was Brophy and Evertson (Note 6), from which several process-outcome relationships were derived. Also used as sources were Blank (1973), the Southwest Education Development Laboratory (1973), and Kounin (1970).

Treatment content. The treatment consisted of 22 principles of small-group instruction, organized under two main headings: management of the entire group, and responding to individual student's answers. Within each major heading, principles were organized by teacher tasks (e.g., getting and maintaining attention, selecting students to answer questions, providing feedback to incorrect answers and failures to respond, and giving praise and criticism). Examples of specific suggestions under these headings were: use a standard and predictable signal to begin transitions; arrange the group to facilitate teacher monitoring and minimize student distraction; call on students in order around the circle rather than selecting them randomly or relying on volunteers; use sustaining feedback (simplifying questions) after student errors when appropriate to the pace; and be specific when giving praise and criticism. Overall, the treatment emphasized maintaining student attention, sequencing information clearly for students, and providing information about the relevant aspects of a question or answer.

Treatment administration. The treatment was described in a short manual that was given to each teacher in the treatment groups. The experimenters met with the teachers to explain the study and leave the booklet, and then returned in about a week to discuss the treatment and answer questions. Control group teachers received no materials.

Subjects. Twenty-seven female first-grade teachers in nine schools agreed to participate after the study was explained to them. All taught students from primarily middle-class, Anglo neighborhoods in a metropolitan school district. The teachers' experience ranged from one to 25 years. All teachers delivered a large part of their reading instruction in small groups.

Design. Before approaching the teachers, three groups of three schools were formed to represent similar SES levels. The three schools in each SES level were randomly assigned to treatment groups. Within a school, all participating teachers were in the same experimental group, so that control teachers did not have direct access to the treatment. Ten of the teachers served as the control group; they had been told that the purpose of the study was to learn what teaching practices related to reading achievement gain, and they did not know that another group was receiving a treatment. Seventeen teachers received the treatment; ten were observed on a regular basis, while seven were not observed. (The latter group was included to assess the impact of observation on treatment effects.) The teachers in the control group and in the treatment-observed group (total observed $N = 20$) were seen once a week between November and May.

Measures. Classroom observation data were collected with a coding system that was directly keyed to the instructional model. Students were tested in May with the Metropolitan Achievement Tests, and their previous

September Metropolitan Readiness Test scores were used as covariables. There were no measures of entering teacher characteristics.

Results: implementation. In general, the treatment teachers had higher implementation scores than the control teachers. However, implementation by the treatment teachers was not consistent across the treatment components; instead, those variables that were most specifically defined in the treatment were most easily implemented by the teachers (e.g., methods for selecting students to answer, giving simplifying feedback after errors).

Comparison of treatment and control groups also revealed that other differences were apparent, especially in the curriculum series used and in the amount of content covered during the year (as measured by the number of basals read). Some of the extraneous group differences may be accounted for by indirect effects of treatment group membership (such as heightened expectations for producing student achievement, which might lead to greater content coverage), but these group differences cannot be accounted for by the specific suggestions in the treatment. Other group differences, such as the primary basal reader used, probably were related to school membership, which was confounded with treatment group assignment.

Results: effects on achievement. There was a significant difference favoring the treatment groups on adjusted achievement. Both treatment groups (observed and unobserved) had higher mean adjusted reading achievement scores than the control group. There were no differences between the two treatment groups, indicating that any treatment effect on achievement apparently was not moderated by observation.

Results: process-outcome relationships. Many of the process-outcome relationships were those expected on the basis of earlier research, and therefore many of the principles in the instructional model were substantiated

by the data in this study. In general, those processes that were associated with reading achievement were also those for which differences in implementation were revealed. That is, treatment group teachers were indeed acting more in line with the treatment in ways that were related to greater achievement. This suggests that the content of the treatment was at least partly responsible for the treatment groups' superior achievement scores.

Conclusions. The treatment was related to changes in treatment teachers' behaviors and to corresponding gains in achievement. However, data also revealed other differences between the treatment and control groups that could not be closely tied to the content of the treatment and which also may have influenced achievement (such as amount of content covered).

Discussion

In all three studies, changes in teachers' behaviors and corresponding changes in student achievement followed the application of a relatively simple treatment. (At least, the treatments were simple in terms of administration.) The purposes of these field-based experimental studies were at least twofold: first, to confirm earlier relationships through an experimental design that allowed some conclusions about causality, and second, to determine if suggestions based on descriptions of effective teaching are sufficient to influence teachers' behaviors.

The second objective was fulfilled to some extent in each study. Although none of the treatments was completely implemented, each was utilized in some way by the treatment teachers.

The first objective, determining causality, is harder to evaluate. In the Missouri Mathematics Effectiveness Project and the Texas First-grade Reading Group Study, there was clear replication of some earlier process-product relationships suggesting that the content of the treatment

contributed to the overall treatment effect. That is, not only were the treatment teachers acting more in line with most parts of the treatment, but the extent to which teachers did so was related to the degree of achievement gain. In the study conducted by the Stanford Program for Teacher Effectiveness, there was no strong evidence of replication of earlier relationships, although there were effects that were associated with the treatment (e.g., implementation of treatment, achievement gains). This discrepency prevents simple conclusions about the content of the treatment causing the associated effects, since, as Cook and Campbell (1976) pointed out, one necessary condition for causal inference is demonstrated covariance of treatments with observed effects. Although the treatment teachers in the Stanford study were behaving in several ways specified by the treatment, there were not strong relationships between those behaviors and student achievement. (There were trends in the expected direction in some cases, but these seldom reached significance.)

Conclusions about causality are somewhat more reasonable in the other two studies, where the treatment teachers were behaving in ways that were associated with student gains. However, in the Texas study, the potential influences of school-related effects on some classroom processes (such as content pacing and choice of materials) can not be eliminated as a possible contributor to the treatment effect. Also, there was no control for a Hawthorne effect, and so there was no way to completely separate effects due to treatment content from effects associated with membership in a treatment group.

The Missouri study is easier to evaluate due to its design. There is less reason to suspect school-related factors as contributors to treatment effects, because 27 schools were sampled. Even though schools were the unit

of assignment, this large number insured that random assignment probably distributed school-related factors across treatment groups. The treatment of the control group and its associated achievement gains made it possible to demonstrate additional effects due to treatment content. Of course, it is not possible to know that the motivation of the control group exactly matched that of the treatment group. On the whole, however, the results of the Missouri study offer the clearest evidence that student learning gains were indeed caused, at least in part, by the content of the treatment.

Any field-based experiment will be open to questions about causal inferences, since there is no way to adequately control all influential factors. However, despite such questions, the three studies did accomplish important objectives. Their overall pattern of results demonstrates the efficacy of process-outcome research in the classroom, and the utility of experimental studies that attempt to modify teachers' behaviors while substantiating earlier research findings. Such studies represent an important next step for researchers of teaching, although there are several issues that must be considered in their design if experimental studies are to answer more questions than they raise.

As noted in the introduction, the dilemma facing the classroom researcher who wishes to conduct experimental studies is to reach a compromise between rigor and relevance. The experimenter must maintain "ecological validity," but must also maintain control of or account for many of the factors that affect teacher behaviors and student outcome. The experimenter must pursue the goal of objective research, while at the same time recognizing the complexities of classroom life and modifying commitments to measures and designs that are more suitable for more controllable conditions.

Several suggestions have been derived from the three studies and have been grouped under two headings: developing treatment content that maximizes the chance for ecologically valid results, and maintaining experimental control in order to isolate effects of treatment content. Many of these suggestions are taken from Brophy (Note 13). Also see Crawford, Gage, and Stallings (Note 14) for other reflections on design. Many of the problems and suggestions exemplify application in a classroom setting of principles first raised by Bracht and Glass (1968), Campbell and Stanley (1963), Cook and Campbell (1976), and Snow (1974).

Developing Treatment Content that Maximizes the Chance for Ecologically Valid Results

The first step in conducting an experimental study of teaching is to decide on the content of the treatment. If the treatment is not chosen wisely, then the experiment, no matter how well designed and conducted, will not produce educationally significant results. Yet, most guidelines on conducting educational experiments (such as Bracht & Glass, 1968; Campbell & Stanley, 1963; Cook & Campbell, 1976) do not devote much, if any, attention to the selection and development of treatments.

Developing a classroom treatment based on research requires that the experimenter do more than simply list research findings. A 'translation' must take place, and this translation involves more than rewording variables to eliminate jargon. The experimenter must organize the research into structures and concepts that are meaningful to teachers in classrooms--that reflect their understanding and knowledge of classroom life.

The three treatments discussed here shared several characteristics that increased their potential for successful implementation and impact. First, the content was relevant to teachers, providing suggestions for dealing with

already existing classroom demands. Second, many of the suggestions were specific. Third, many of the suggestions were organized into clusters with rationales that related them to important goals.

Selecting relevant content

The treatments in the three studies may be viewed as collections of strategies and techniques that are applicable to common demands of the classroom. They reflected the complexity of the teachers' work in that each treatment was presented as a system of instruction that was organized around several routine tasks of teaching the basic curriculum. For example, the Missouri study emphasized presentation of new information, monitoring student practice, and accounting for students' work. The Stanford study provided suggestions for behavior control, arranging students for instruction, and questioning students during a discussion. The Texas study focused on group-management techniques and questioning and feedback strategies. All of these categories of behaviors describe the tasks of all teachers in comparable circumstances; the treatments simply focused the teachers' attention on the tasks and recommended ways of performing them more efficiently or in a more goal-directed manner. The teachers were not asked to perform additional or unusual tasks.

The ultimate goal of each treatment was improved achievement in basic skills through improvement of daily instruction--a goal that is likely to be shared by most teachers. Therefore, the studies can be contrasted with others in which the goals may not be shared by the teacher (e.g., to validate some premise in a theoretical model) or that require additional work by the teachers (e.g., examining ways of presenting the content of an experimental science curriculum).

Teachers' responsibilities are vast, and there is competition for teacher attention and energy (Doyle, 1977; Jackson, 1968). The results of these three studies indicate that research findings can claim their share of teacher attention when they represent useful information that assists teachers in fulfilling their major instructional goals.

This does not mean that teachers will not change in more drastic ways, given appropriate institutional support. The point made here is that simply providing relevant information to teachers, without providing additional support or removing demands, was sufficient to yield different, and presumably better, teaching. Some probable reasons are the practicality of the content and the ease of incorporating it into a daily routine.

Specificity of Treatment

In each study, the principles in treatments that were most clearly implemented by the treatment groups were those that were most specifically defined. That is, they were presented in behavioral terms--what the teacher was to do. Principles that were less well implemented tended to be more general, on a higher level of abstraction or complexity, so that it was less clear to the teachers exactly what behaviors were expected. (Of course, it may also have been the case that the measures were less clearly defined and subject to more error.) For example, in the Missouri study, suggestions about conducting review sessions and checking homework were implemented by the treatment teachers at a higher level than control teachers. However, recommendations for emphasizing meaning and understanding in the development portion of the lesson were not implemented by the treatment teachers. In the Stanford study, principles describing selection of students to answer questions were well implemented, while principles describing the need to accurately target discipline corrections did not lead to higher levels of

appropriate teacher behaviors. In the Texas study, treatment teachers implemented suggestions about selecting students and giving feedback to their answers, but they did not utilize suggestions about using students as peer models or breaking the group when students revealed different rates of learning.

It can be concluded that the kind of treatment offered in these three studies (i.e., relatively minimal in terms of efforts required to present it and learn it) will have its strongest impact on teacher behaviors that are clearly and simply described. More complex teaching behaviors may require more extensive treatment than was offered in these three studies. The content and the methods of presenting a treatment are related; this point must be considered by researchers who wish to provide treatments and assess their impact on behaviors.

These points are comparable to those made by Doyle and Ponder (1977), who suggested that the "practicality ethic" determines whether or not teachers will effectively use advice. Practicality depends on three qualities: operationality (easily translated into behavior); congruence with the teacher's own role definition; and efficiency in terms of the teacher's cost and time. These three suggestions imply that a treatment will be implemented most easily when it is specific in terms of routine teacher behaviors, and when it provides a rationale that effectively relates the behaviors to the teacher's goals of instructing the students. Also, the behaviors must not make extensive demands on the teacher's time and energy, or at least none that are not compensated for in some way.

Clustering of Specific Suggestions

However, in requiring that a treatment be operational and relevant in terms of routine classroom tasks, a problem arises. The treatment must be

specific enough to allow translation into actual behavior, but there are inherent difficulties with very specific advice. Specific suggestions must be imbedded within larger principles, since no isolated behavior can be appropriate all of the time. It is not possible or desirable to list all possible situations and the variables defining appropriateness. Such an approach implies that teachers memorize techniques and apply them arbitrarily.

A more reasonable approach to treatment design is to identify general principles of effective instruction and to cluster specific strategies under each. The specific strategies then serve as examples of the larger principle so that teachers are provided with operational concepts. When treatments are communicated in this way, the suggestions about particular techniques are imbedded in contexts and supported by rationales that prevent the treatment from appearing arbitrary and inflexible.

The use of meaningful clusters was present in each of the three studies reviewed here, although they were different in each study. In the Missouri study, sequence within the lesson was the basis of the treatment components, and objectives for each part of the lesson were presented as the guiding concepts. Since this study represented a nearly complete treatment for a particular setting, it was possible to use temporal sequence as a consistent organizer. The Stanford study used three broad aspects of teaching to organize the specific suggestions: behavior control and management, instructional methods, and questioning and feedback. Since the treatment in the Stanford study included only variables from earlier studies that yielded significant correlations, their clusters were primarily based on the variables available. In the Texas study, a combination of lesson components (getting

attention, introducing the lesson), and teaching tasks (calling on students, giving feedback) were used to group the variables.

Each treatment included some discussion of contextual distinctions that should be considered when applying the principles. In the Missouri math study, teachers were urged to consider student understanding at various stages in the lesson, and to select their next steps according to that diagnosis. In the Stanford study, teachers were given different guidelines for more and less academically oriented students. In the Texas study, lesson pace and type of question were to be taken into account in responding to different types of students answers.

The purpose of presenting suggestions within clusters and with contextual qualifications is to help teachers choose among alternative strategies according to a rationale that defines appropriateness. That is, the purpose of "treating" teachers is not to encourage them to use a specific technique every time it is possible, but instead to optimize use of several techniques or strategies--using them when appropriate and avoiding their use when inappropriate. Training teachers to make such decisions requires a conceptual framework with which they may examine their classroom tasks along with a rationale that explains why certain specific strategies are or are not effective in various contexts.

The implications of this for treatment design are that specific suggestions must be placed within a meaningful framework. This is an important part of meeting the first objective of classroom experimental design: maintaining ecological validity. Teachers daily make thousands of decisions about their teaching tasks and no treatment, no matter how thorough, will replace the teacher's use of his or her own judgment of what is best to do at any given moment.

One important implication of this for the experimenter who designs a treatment is that one must often interpret beyond the data available from the original sources, although caution against overinterpreting must be observed. The original correlational research has most likely been reported in terms of variables derived from an observation instrument. The experimenter must decide how literally to present those measures, and how much interpreting is reasonable in order to accurately depict the results. Results must be considered as more than the specific measurements, and presented within a context of meaning that provides rationale and purpose. That is, a measure of a behavior in one setting does not translate directly into recommendations for practice in that setting or in any other. Behaviors are meaningful only when considered within a larger context. It is short-sighted to consider any behavior that is correlated with outcome as necessarily having a direct connection to outcome, whether causal or otherwise. In many cases, it is more likely that the behavior is part of a sequence of events that are all associated with outcome, and another precursory event may have a much more direct connection to the outcome as well as to other intermediate behaviors.

This is, of course, the rationale behind various statistical approaches to grouping variables. The point here is that it must also be done on a conceptual level in defining specific behaviors to incorporate in a treatment, even if some intuitive leaps must be made. There are many possible interpretations of the functional meaning of any variable, and these interpretations affect the presentation of results in a treatment. The experimenter who is familiar with classroom life and teachers' responsibilities is likely to be more successful at making valid and realistic interpretations than a researcher who has not worked extensively with teachers and classroom data.

As an example, consider two alternatives to presenting the findings from several studies that more effective teachers (i.e., whose students have higher adjusted achievement) spend relatively more time on academic content and relatively less time on procedural, organizational, and behavioral matters. If one were to go directly from the variable to the treatment, the recommendation would be to reduce time spent in nonacademic matters. However, this ignores the fact that such time allocation may well be an outcome of other, more basic managerial practices. As such, too much time spent on procedural matters is a symptom of other problems, and treating the symptom may not directly change the underlying cause. Therefore, one might present the behavior as an indication of poor time management with suggestions for direct treatment of the probable cause. Of course, this is likely to be a less data-based approach, probably requiring some supplemental information to data found in the correlational studies. The experimenter's dilemma, then, is juggling common-sense requirements with objective data in creating a treatment with a basis in earlier research that is also likely to be relevant and meaningful to teachers.

Another way in which "common-sense" enters into treatment development is the recognition that changes in context or setting may change the appropriateness of a suggestion. For example, there are times when more effective teachers devote more time, relatively, to procedural and organizational matters, such as the beginning of the school year or when introducing major organizational changes (Emmer, Evertson, & Anderson, Note 15). Obviously, the dimension that defines effectiveness is "appropriateness for a given set of circumstances." Although one's correlational data source may not yield these contextual distinctions, the credibility of a treatment (and hence, the likelihood of implementation) may be enhanced by including

qualifications that are based on "common sense" and a knowledge of classroom ecology.

Maintaining Experimental Control to Isolate Effects of Treatment Content

If adequate experimental control is not maintained, the treatment, no matter how well conceived and applied, cannot be evaluated completely. In the traditional sense, "experimental control" implies the elimination of any contaminating or modifying influences by carefully matching groups on all but the independent variables or by systematically varying other factors. Hence, the treatment is the only possible explanation for differences between experimental and control groups. It is obviously not desirable to exert control in this sense in classroom studies (since this would limit generalizability, which usually has high priority). However, it is important to be able to isolate effects due to the treatment and those due to other factors that are unrelated or indirectly related to the treatment content or process. Control in classroom experimental studies may be exerted both through initial assignment of classes to treatment groups, and through ongoing measurement of classroom processes. The latter may be used to statistically control for some factors, and may also be used for descriptive purposes to aid in interpretation. The purpose of such ongoing measurement is to learn as much as possible about what actually occurred in the classrooms during the study.

The results of the three investigations suggest several ways that experimental classroom studies may be carefully monitored. Many of the suggestions involve choice of measures, while others reflect concerns with design.

Selecting and Using Measures in Experimental Classroom Studies

Implementation measures. One purpose of this type of study is to determine the effects of the treatment on teacher and student behavior. In particular, one important question is how the teacher changed his or her behaviors as a result of the treatment. Therefore, the observation instrument used should be targeted to the variables of most interest--the use (and possible misuse) of the principles in the treatment. When the treatment is based primarily on one other study, the observation instrument used previously may be useful in the treatment study (as was true in the Missouri study). However, in most cases, a new set of classroom measures must be developed or the resulting data will not adequately answer the questions about the treatment. This was one problem encountered by the Stanford experimenters, who selected as the primary instrument a system used in one of their four source studies. Variables from the other three studies could not be clearly mapped onto the instrument, so that parts of the treatment derived from those other studies could not be evaluated completely. The Texas study is an example of use of a "content-referenced" coding system, since it was developed especially to test the hypotheses raised in that particular study. Therefore, each element of the treatment was clearly reflected in the measures.

After choosing or developing a relevant set of measures, they should be used to gather sufficient information to evaluate implementation in both the treatment and the control classes. It is not unlikely that there will be some natural use of the treatment principles in control classes, since the research on which the treatment is based was conducted in classrooms where teachers were using the skills. In fact, it is quite possible that some of the teachers in the control group will be "implementing" part of the treatment at as high or higher a level than some treatment teachers. Therefore, in order

to accurately assess the effects of treatment content, it will be necessary to separate users from non-users. Research on the implementation process (Hall & Loucks, 1977) has demonstrated that an educational innovation can be most effectively evaluated when actual levels of use are considered rather than arbitrary distinctions based on treatment group membership.

Short-term outcomes. Adequate measures of short-term outcomes (through an observation system that allows recording of events in sequence) will allow the experimenter to more accurately assess the treatment effects in terms of routine classroom events (e.g., student behavior following teacher corrections, student answers to a teacher question, and students' smoothness in following a new procedure). Therefore, the final criterion of effectiveness of treatment does not have to be limited to end-of-the-year achievement and/or attitude. Even if end-of-the-year measures are desirable, inclusion of the short-term outcomes will clarify causal connections between the treatment variables and the eventual outcome. Short-term outcomes may also make apparent any connections between the treatment and outcomes that were not expected.

For example, in the Texas study, the treatment suggested that transitions could be made more effective by using standard signals to notify the students. Most teachers, including control teachers, did this, and so there were no group differences in implementation of that principle. Neither was there a relationship with outcome due to the restricted variance. (This did not indicate that the principle was invalid, but that it represented a very common behavior.) However, there were differences between the groups on measures of the efficiency of transitions (with the treatment group having shorter, smoother ones). Since there were no measures of preexisting differences, it is possible that the treatment teachers had smoother transitions to begin

with. However, an alternative explanation is that the treatment may have caused the teachers to focus on their transitions and work harder to make them more efficient, even though the strategies given in the treatment (i.e., standard signals) were obviously not ones that made a difference. However, the short-term outcome measure (efficiency of transitions) provided support for an alternative explanation of treatment effects: increasing awareness of a problem area leads to improved behavior, regardless of specific suggestions.

Pretreatment measures of classroom processes. Before the treatment is administered, there should be some measures taken of classroom processes that are expected to change over time due to the treatment. Comparable measures should be taken in both control and treatment classrooms. These scores may then be used as baseline measures in order to compare changes as a result of the treatment, and to demonstrate how comparable the treatment and control groups were initially. Depending on the research question, the pretreatment measures could be used to assign teachers to conditions, if it were desirable to have treatment and control groups initially comparable on some process measures. For example, a study aimed at improving teachers' classroom management strategies might initally draw a sample of teachers at several levels of proficiency, ranging from teachers who are already excellent managers to teachers who need a few suggestions in a few areas, to teachers who need to improve in many ways. Depending on the purpose of the study, the experimenters could use baseline information to form matched treatment groups, or to eliminate some teachers from the subject pool so that the study included only one level of entering proficiency.

The Stanford study was the only one of the three to include pretreatment measures, and thus it was possible to examine treatment effects over time, and

to demonstrate several differences between the treatment and control groups before the treatment. Using baseline measures as a covariate in comparisons of the groups prevented some false conclusions about treatment effect. The other two studies had to rely on comparisons of group means, without knowing how many differences were present before the treatment. One must assume that random assignment prevents major differences, but because schools were the unit of assignment, group means on classroom process measures may have been influenced by school membership, especially in the Texas study, where observations were made in only six schools.

Pretreatment measures of teacher, student, and school characteristics. In addition to measuring classroom processes that will be used to evaluate the treatment effect, it would be wise to obtain other pretreatment measures of teacher, student, and school characteristics, to be used either as covariates in the analyses of treatment effects, or to use to balance experimental groups from the beginning. The list of possible characteristics is infinite, and the choice of measures to be taken in any particular study is a function of the research questions. However, some general types of measures will likely be appropriate in almost any study. Student entering achievement should probably be used as one way of balancing groups, as was done in the Stanford study. If this is not possible, then entering achievement should be obtained and used in analyses of treatment effects, as was done in all three studies. Other student characteristics may also be important, such as motivation.

Teacher characteristics were measured in two of the studies, and were shown to interact with or affect the influence of the treatment. In the Stanford study, teacher verbal fluency and pretreatment attitudes toward structuredness were important predictors of later behaviors; unfortunately, these were confounded with treatment group membership, and the treatment

effect could not be completely separated from them. This suggests the importance of measuring teacher attitudes and characteristics that may influence teachers' receptiveness to a treatment (i.e., it is easier to implement principles that are in accordance with one's own values and beliefs). The Missouri Math Study data supported this by yielding complex interactions between teacher types, student types, and treatment group membership (Ebmeier & Good, 1979).

School characteristics should also be taken into account, especially if the experimenter chooses to make treatment assignments within schools. Even if it is possible to have only one teacher per school involved, there are still factors related to the school environment that should be considered. For example, in many schools where teachers team for instruction in some subjects, the amount of time and the pace at which the curriculum is covered is determined by school policies. Instructional resources, both material and human, will vary from school to school. All of these factors may affect the quality of instruction received by the students, and may obscure effects of the treatment.

Measuring other setting characteristics. In addition to measuring process variables that demonstrate use of the treatment and other factors that may moderate the effects of the treatment, the experimenter should also measure as many other processes and setting characteristics as possible. Needless to say, this number will be limited, of practical necessity. Probably the most desirable type of additional information would be a few instances of rich descriptions that include other aspects of the classroom and other teacher and student behaviors than those focused on in the treatment. Using such descriptive information, it would be possible to examine the treatment variables in context. Quantitative, low-inference coding systems

often decontextualize variables in order to compare them across several classrooms. This is necessary, of course, for some analyses, but when unexpected results arise, it is often difficult to evaluate them with only coded data. As an example of how such descriptive data could have added to the information provided by an experimental study, consider the Stanford study's variable of "teacher movement." In the Stanford treatment, teachers were urged to move around the room a lot in order to monitor better. This principle was supported by earlier research that positively related teacher movement to achievement. However, when process-product correlations were computed for the sample in the Stanford study, there was a significant negative correlation with achievement. This could have been due to chance variation, but another possible explanation presents itself, although it cannot be confirmed. Several of the schools in the study were involved in a state-supported program that encouraged individualized assignments and small-group work. It is possible that this represented a major change for many of the teachers, if they were accustomed to more traditional instructional patterns. Teacher movement in the context of an unfamiliar instructional format may well represent a different functional behavior from movement in a familiar format. In a new situation, movement might represent lack of organization, and excessive motion in order to keep up with what is going on in the room; that is, the teacher may be pulled about, so that the movement is more a reaction to immediate needs than it is a planned approach to monitoring. Within a more familiar setting, teacher movement might represent purposeful monitoring, so that movement is appropriate to larger instructional goals. However, a standardized measure of teacher movement might well obscure the dimension of goal-directedness. If enough teachers in the Stanford study were actually engaged in less appropriate movement due to

new demands, then it is not surprising that a negative correlation with achievement was obtained. In the earlier studies, movement might have been more appropriate in the observed classes, thus the positive relationship with achievement. The point is that one cannot equate functional meaning of classroom occurrences by using comparable measures if those measures do not incorporate dimensions of appropriateness for the setting and the goals of that setting. This does not mean that quantitative, coded measures do not have their place in such an investigation; it is important that some standard measures be taken in each class. However, they should be supplemented with other information, either through descriptive records or through more complex coding that would yield more complete descriptions of the setting.

Experimental Design to Isolate Treatment Effects

The question of how best to design an experimental study involves more than the selection of instruments and variables. A more traditional concern is the creation of groups of individuals who receive comparable versions of the treatment or who serve as controls. One dilemma faced in any treatment study is to separate the effects of the content of the treatment per se from incidental effects of assignment to a group that is getting some special treatment. Special attention, regardless of the nature of that attention, may be sufficient to lead to changed behavior. Likewise, being a member of a control group and knowing that others are receiving special attention and help may have unintended effects. In either case, depending on the subjects' perceptions and attitudes, the effects could be either improved or worsened performance that is not related to the content of the treatment.

In designing an experimental study of teaching, it is important to keep such potential effects in mind and to prevent or at least to monitor them. One way is that utilized in the Missouri study: deliberately trying to

heighten motivation in the "control" group (actually a delayed-treatment group) so that they would be more like the treatment group in that respect. The achievement gains of the Missouri study's delayed-treatment group confirm the importance of what was probaly an increase in personal efficacy. Therefore, the extent to which all teachers are motivated to perform well should be controlled or monitored.

There is also evidence that making teachers more aware of their behaviors, even without offering suggestions for changing behavior, may often lead to desired results. Good and Brophy (1974) gave objective feedback to teachers on the frequency and quality of contacts they had with different students in the room. Without any suggestions about changing their patterns of interaction, the teachers demonstrated more equal treatment of students on such variables as calling on students to answer questions and extending their opportunities to interact with the teacher after an initial error. (These are two areas where students who inspire "low expections" may be short-changed.) Feedback alone was sufficient to heighten the teachers' awareness, and that was followed by changed behavior in a desirable direction.

These two points--motivation to do well and increased awareness of critical aspects of teaching--may be important effects of treatment group membership. If the purpose of conducting an experiment is to examine the relationship of specific teaching styles and behaviors to outcomes, then motivation to teach well and awareness of the topics in the treatment should be monitored, perhaps by creation of several control groups or alternative-treatment groups. In such a case, comparison groups assume a role beyond that of representing untreated controls. They allow the experimenter to identify exactly which aspects of a treatment are influential in changing teaching behaviors. Therefore, one can improve construct validity, as

described by Cook and Campbell (1976), who also recommended multiple treatment groups.

For example, the Stanford study included two treatment groups that varied the extent of training. One group received written materials while the other also had training sessions. In the Texas study, one treatment group had regular observations while another was not observed.

Other dimensions of treatment administration that could be varied within an experimental design include the specificity level of the treatment and the amount of individualized feedback given to teachers. One might hypothesize that these treatment characterisitics would interact with teacher experience and level of expertise.

Conclusion

Obviously, it is impossible to incorporate all of these points into a single study unless one has extensive resources. They have been presented as a synthesis of current knowledge about and experience with the conduct of large-scale experiments in natural classroom settings.

As more pressure is put on researchers in the social sciences to apply their knowledge to the resolution of humanity's problems, the use of field settings will increase as greater relevance and applicability are sought. The use of the experimental paradigm in such settings holds much promise as a method geared both to practical problems and to scientific speculation. It is essential that researchers not lose sight of the two requirements that field settings force upon them: maintaining experimental control through careful monitoring of many factors, and knowing as much as possible about the natural settings. This latter requirement implies much about the first: one must understand what phenomena are most salient and relevant before selecting measures and experimental treatments.

Although the reviewed studies were conducted by persons who consider themselves primarily as researchers of teaching rather than as teacher educators, the results and suggestions derived from the studies have implications for the more general study of teacher education. Research on teacher education and staff development has generated a great deal of interest of late, and many questions have been raised about how to pursue such research. All of the problems faced in conducting experimental studies of classroom teaching are present and even magnified when one attempts to examine the effects of programs of teacher education on both teachers and students. As was true with the experimental studies of teaching, a major difficulty is the need to consider simultaneously questions of content selection (i.e., what are the effects of instruction in content about various components of teaching?) and process (e.g., what are the effects of various ways of communicating that content?) In addition, there are many potential intervening variables. However, these results suggest that such questions are researchable. Many of the points raised in this article about selection of treatments, development of treatment groups, and choice of measures to monitor classroom occurrences may be adapted to the study of teacher education programs.

This review has focused on the elementary classroom as an example of a natural setting in which complex but detailed experiments have been successfully conducted. As knowledge about classrooms and teaching effects accumulates, it is likely that more experimental studies such as these will be conducted. If these three studies can be considered portents of the future, researchers may soon be able to point to several significant examples of their beneficial influence on educational practice, while at the same time they have

refined their knowledge of classroom environments and the processes of teacher training.

Reference Notes

1. Medley, D. Teacher competence and teacher effectiveness: A review of of process-product research. Washington, D. C.: American Association of Colleges for Teacher Education, 1977.
2. Program on Teaching Effectiveness. An experiment on teacher effectiveness and parent-assisted instruction in the third grade. Stanford, Cal.: Stanford University, Center for Educational Research at Stanford, 1978.
3. Stallings, J. Teaching basic reading skills in secondary schools: a second year study. Paper presented at annual meeting of American Educational Research Association, San Francisco, April, 1979.
4. Koehler, V. Methodology for research on teacher training. Paper presented at the conference, "Exploring Issues in Teacher Education: Questions for Future Research," University of Texas, Austin, Texas, January, 1979.
5. Program on Teaching Effectiveness. A factorially designed experiment on teacher structuring, soliciting, and reacting (R&D Memorandum No. 147). Stanford, Cal.: Stanford Center for Research and Development in Teaching, 1976.
6. Brophy, J., & Evertson, C. Process-product correlations in the Texas Teacher Effectiveness Study: Final Report (R&D Rep. No. 4004). Austin, Texas: University of Texas, Research and Development Center for Teacher Education, 1974.
7. McDonald, R., & Elias, P. The effects of teaching performance on pupil learning, Beginning Teacher Evaluation Study: Phase II, Final Report: Vol. I. Princeton, N. J.: Educational Testing Service, 1976.

8. Soar, R. Follow Through classroom process measurement and pupil growth (1970-1971), final report. Gainesville, Florida: University of Florida, Institute for the Development of Human Resources, 1973.

9. Stallings, J., & Kaskowitz, D. Follow Through classroom observation evaluation, 1972-73. Menlo Park, California: Stanford Research Institute, 1974.

10. Mitman, A. Teacher self-report in an experiment on teacher education and effectiveness. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

11. Crawford, J., & Stallings, J. Experimental effects of in-service teacher training derived from process-product correlations in the primary grades. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

12. Stayrook, N., & Crawford, J. An experiment on teacher effectiveness and parent-assisted instruction in the third grade: The observational data. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

13. Brophy, J. Training teachers in experiments: Considerations related to nonlinearity and context effects. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

14. Crawford, J., Gage, N., & Stallings, J. Methods for maximizing the validity of experiments on teaching. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

15. Emmer, E., Evertson, C., & Anderson, L. The first weeks of class . . . and the rest of the year. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

References

Anderson, L., Evertson, C., & Brophy, J. An experimental study of effective teaching in first-grade reading groups. Elementary School Journal, 1979, 79, 193-223.

Blank, M. Teaching learning in the preschool: A dialogue approach. Columbus, Ohio: Charles E. Merril, 1973.

Borich, G. (Ed.). The appraisal of teaching: Concepts and processes. Reading, Massachusetts: Addison-Wesley, 1977.

Bracht, G., & Glass, G. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.

Brophy, J. Teacher behavior and its effects. Journal of Educational Psychology, in press.

Brophy, J., & Evertson, C. Learing from teaching: A developmental perspective. Boston: Allyn & Bacon, Inc., 1976.

Brunswik, E. Perception and the representative design of psychological experiments. Berkeley, Calif.: University of California Press, 1956.

Campbell, D., & Stanley, J. Experimental and quasi-experimental designs for research on teaching. In N. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally & Co., 1963.

Cook, T., & Campbell, D. The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally & Co., 1976.

Doyle, W. Learning the classroom environment: An ecological analysis. Journal of Teacher Education, 1977, 28, 51-55.

Doyle, W., & Ponder, G. The practicality ethic in teacher decision-making. Interchange, 1977, 8, 1-12.

Dunkin, M., & Biddle, B. The study of teaching. New York: Holt, Rinehart, & Winston, 1974.

Ebmeier, H., & Good, T. An investigation of the interactive effects among student types, teacher types, and instruction types on the mathematics achievement of fourth-grade students. American Educational Research Journal, 1979, 16, 1-16.

Good, T. Teacher effectiveness in the elementary school. Journal of Teacher Education, 1979, 30, 52-64.

Good, T., Biddle, B., & Brophy, J. Teachers make a difference. New York: Holt, Rinehart, & Winston, 1975.

Good, T., & Brophy, J. Changing teacher and student behavior: An empirical investigation. Journal of Educational Psychology, 1974, 66, 390-405.

Good, T., & Grouws, D. Teaching effects: A process-product study in fourth-grade mathematics classes. Journal of Teacher Education, 1977, 28, 49-54.

Good, T., & Grouws, D. The Missouri Mathematics Effectiveness Project: An experimental study in fourth-grade classrooms. Journal of Educational Psychology, in press.

Hall, G., & Loucks, S. A developmental model for determining whether the treatment is actually implemented. American Educational Research Journal, 1977, 14, 263-276.

Heath, R., & Nielson, M. The research basis for performance-based teacher education. Review of Educational Research, 1974, 44, 463-484.

Jackson, P. Life in classrooms. New York: Holt, Rinehart, & Winston, 1968.

Kounin, J. Discipline and group management in classrooms. New York: Holt, Rinehart, & Winston, 1970.

Rosenshine, B. Classroom instruction. In N. Gage (Ed.), The psychology of teaching methods, seventy-fifth yearbook of the National Society for the Study of Education (Part I). Chicago: The University of Chicago Press, 1976.

Rosenshine, B., & Furst, N. The use of direct observation to study teaching. In R. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally & Co., 1973.

Southwest Educational Development Laboratory. Bilingual kindergarten program, in-service manual (Vol. I). Austin, Texas: National Educational Laboratory Publishers, Inc., 1973.

Shulman, L. Reconstruction of educational research. Review of Educational Research, 1970, 40, 371-393.

Snow, R. Representative and quasi-representative designs for research on teaching. Review of Educational Research, 1974, 44, 265-291.

Footnotes

[1]A fourth experimental study at the high school level has been performed by Stallings (Note 3). Because of the different target population and differences in treatment design (she had more interaction with individual teachers), it has not been included here.